改进的高效模糊 C 均值聚类算法

丁旭晨，林锦贤
（福州大学 数学与计算机科学学院，福建 福州 350108）

摘 要：在数据采集过程中结合网格聚类算法提高计算效率，为了保存采样数据的分布特点引入权值。根据类别中心密度高、权值大的特征采用寻找连通分量的方法初步确定聚类中心，在此基础上结合自适应免疫算法，动态地确定聚类中心及其类别数。进而使 FCM 算法跳出局部最优，最大可能地得到全局最优解。

关键词：聚类；权值合理性；自适应免疫算法；连通分量

中图分类号：TP311 **文献标识码**：A **文章编号**：1674-7720(2013)23-0074-03

Improved efficient fuzzy C-Means clustering algorithm

Ding Xuchen, Lin Jinxian
(School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: The grid clustering algorithm is combined to improve the computational efficiency in the data collection process, and the right values are introduced to preserve the distribution characteristics of the sampled data. Depending on the characters of large weight and density in the category center, methods of finding connected components are adopted to initially identify cluster centers. On this basis, the adaptive immune algorithm is combined to dynamically determine the cluster centers and the number of categories. Thereby the FCM algorithm is able to escape from local optima, and the maximum possible to get a global optimal solution.

Key words: clustering; weights rationality; adaptive immune algorithm; connected component

FCM 算法是当前最受关注的聚类方法之一。本质上，FCM 算法是一种初始化数据与聚类结果之间的映射方法，其求解过程采用一种爬山技术进行迭代来寻找局部最优解。当初始化完成后，相应的聚类结果就已经确定了。所以 FCM 算法对初始化数据比较敏感，有可能陷入局部最优解，无法得到全局最优解。

针对此问题目前主要的解决方法有两种：(1)在聚类的过程中进行全局随机搜索[1]，Xinchao 等人采用模拟退火算法。对当前聚类所获得的结果采用退火计划进行扰动，并以一定的概率接受扰动后的结果为当前最优解，进而跳出局部最优解。此算法要求足够多的扰动次数，配以严密的扰动计划来获取全局最优解，即此算法要求大量的计算。(2)改善 FCM 算法的初始化条件，选择合适的聚类中心[2]。在初始化问题上，本文受第二种方法启发，把网格划分加权[3]、寻找连通分量、自适应免疫算法[2]相结合。对原始数据进行初始化处理，动态寻找初始类别数和相应的聚类中心。进而跳出局部最优，最大程度地找到全局最优解。本文将原始的 FCM 算法和改进的 FCM 算法进行对比分析。实验表明，该改进方法能够有效地降低时耗，提高聚类收敛速率。

1 权值合理性证明

假设 $S=(S_1,\cdots,S_m)$ 是从网格聚类中获得相应子空间中的 m 个中心点。第 i 个中心点所对应的特征值为：$S_i=(S_{i1},\cdots,S_{in})$，表示有 n 维空间。以下给出几个定义：

定义 1 网格密度。相应网格中数据点的个数与该网格空间大小的比值，计算公式为：

$$\rho_v=\frac{N_v}{R_v} \tag{1}$$

其中 $R_v=\prod_{i=1}^{n}(b_i-a_i)$；$b_i$、$\alpha_i$ 分别是单元格在第 i 维上的左右边界；N_v 为网格中的数据点数。

定义 2 网格聚类中心。设在此网格中有 i 个数值点：$\{v_1,\ldots,v_i\}$，其聚类中心为：

$$C_k=\frac{\sum_{j=1}^{i}V_j}{i} \tag{2}$$

定义 3 网格聚类中心的权重。为了保存原始数据的分布特征，引入了权重：

$$W_i=\frac{\rho_i}{\bar{\rho}} \tag{3}$$

其中 $\bar{\rho}=\frac{\rho_1+\cdots+\rho_m}{m}$。

定义 4 加权合理性证明的 2 个条件：(1)加权后的聚类中心能够有效地偏向于权重比较大的类别。(2)加权后的聚类中心必须要在相应的样本点范围之内，不能偏离出相应的样本点。

定义 5 自适应免疫算法中。变异后的聚类中心定义为：

$$C=C-\alpha(C-X) \tag{4}$$

其中 $a=k\frac{\bar{\rho}}{\rho_i}d$，$d$ 为各样本与聚类中心的距离，k 为常量系数。

FCM 算法是在 k-means 算法的基础上引入模糊概念。在本质上和 k-means 算法是一致的。为了证明的简便，本文是基于 k-means 算法进行证明的。证明过程中，加权之后的样本值为：

$$S=\{\frac{\rho_1S_1}{\bar{\rho}},\frac{\rho_2S_2}{\bar{\rho}},\cdots,\frac{\rho_mS_m}{\bar{\rho}}\}$$

相应的聚类中心为：

$$C_i=\frac{\frac{S_{1i}\rho_1}{\bar{\rho}}+\frac{S_{2i}\rho_2}{\bar{\rho}}+\cdots+\frac{S_{mi}\rho_m}{\bar{\rho}}}{m}=\frac{S_{1i}\rho_1+S_{2i}\rho_2+\cdots+S_{mi}\rho_m}{\rho_1+\rho_2+\cdots+\rho_m}$$

对于证明，本文采用的是数学归纳法：

当 $m=1$ 时，即此时只有一个样本点。

$C_1=S_i\rho_1/\rho_1=S_i$，显然满足合理性条件，成立。

假设当有 $m-1$ 个样本时成立：

$$C_{i(m-1)}=\frac{S_{1i}\rho_1+S_{2i}\rho_2+\cdots+S_{m-1i}\rho_{m-1}}{\rho_1+\rho_2+\cdots+\rho_{m-1}}$$

当有 n 个样本时，对以上等式拆分，同除以 $\rho_1+\rho_2+\ldots+\rho_{n-1}$ 后得到如下式子：

$$C_{im}=\frac{\frac{S_{1i}\rho_1+S_{2i}\rho_2+\cdots+S_{m-1i}\rho_{m-1}}{\rho_1+\rho_2+\cdots+\rho_{m-1}}+\frac{S_{mi}\rho_m}{\rho_1+\rho_2+\cdots+\rho_{m-1}}}{1+\frac{\rho_m}{\rho_1+\rho_2+\cdots+\rho_{m-1}}}$$

$$=\frac{C_{i(m-1)}+\frac{S_{mi}\rho_m}{\rho_1+\rho_2+\cdots+\rho_{m-1}}}{1+\frac{\rho_m}{\rho_1+\rho_2+\cdots+\rho_{m-1}}}$$

取 $\alpha=\frac{\rho_m}{\rho_1+\rho_2+\cdots+\rho_{m-1}}$，$\alpha\in(0,\infty)$ 得 $C_{im}=\frac{C_{i(m-1)}+S_ma}{1+a}$。此式可以看做是前 $m-1$ 个样本合理的聚类中心点和第 m 个样本之间求聚类中心。其中前 $m-1$ 个样本的聚类中心的网格密度为 1，而第 m 个样本的网格密度为 α。

对其进行求导得：$C_{im}'=\frac{S_m+C_{i(m-1)}}{(1+a)^2}$

不失一般性，本文讨论 $S_m>C_{i(m-1)}$，即 C_{im} 为增函数的情况，C_{im} 随着 α 的变大而变大。当 $\alpha\to\infty$ 即 $\rho_m\to\infty$ 时，$C_{im}=S_m$ 达到最大，此时 S_m 是高密度的数据点；当 $\alpha\to0$ 时，($C_{im}=C_{i(m-1)}$)达到最小，此时 $C_{i(m-1)}$是高密度的数据点。所以 C_{im} 满足合理性条件，即权值的取法是合理的。

2 改进的 FCM 算法

2.1 确定初始化聚类中心及类别数

(1)引入网格加权

将 FCM 算法与网格聚类的方法相结合，目的是引入网格聚类的优点：在处理数据时与数据对象的数目无关，只与每维空间所划分的单元数目有关，保证聚类方法的高效性。在 FCM 算法中引入权重，对于网格密度大的空间，可以赋予相应大的权重，如定义 3。最大程度保存原始数据分布特点，保证聚类结果的有效性。

(2)利用权值寻找连通分量法，确定初始化聚类中心

类别中的数据在一定范围内是密集的，即相应的网格数据之间连通，并且相应的密度权值从最高密度中心向外递减，密度权值最大的网格肯定包含于某一类别。步骤如下：

①找到权值最大的网格，递归寻找所有与之连通的网格。递归结束需满足以下条件：没有与之相连通的有数据的网格；虽然存在与之连通的网格，但是该网格的权值大于当前网格的权值。

②把步骤①中找到的连通网格看成一个初始类，并删除选中的节点。然后重复上面的步骤，直到所有的数据网格都被处理完为止。

(3)对初始化聚类中心，与自适应免疫算法[2]进行结合。把聚类中心点看成是抗体，把相应的数据点看成是抗原。该抗体能够识别周围一定范围内的抗原。其中 ε,σ 是预先确定的门限值。具体步骤如下：

①对中心进行加权，权值以此中心所识别的抗原个数为依据。

②遍历所有抗原计算其对于各抗体的亲和度，并根据式(3)进行相应的变异。

变异率与抗原与抗体之间的距离成正比，与抗原的权值成反比，与亲和力成反比，根据式(4)进行计算。

③把抗体之间的距离小于 ε 的抗原进行合并，形成记忆细胞。

④重复②、③直到无满足条件的操作为止。

⑤去除不同记忆细胞中距离小于 σ 的个体，即对其进行合并。所形成的新记忆细胞就是最后的初始化聚类中心，记忆细胞的个数就是初始化的类别数。

2.2 FCM 算法执行过程步骤

(1)样本的归一化处理。

(2) 使用改进的初始化方法确定类别数和相应的聚类中心。

(3)根据式(5)和式(6)[3]确定隶属度矩阵 $\boldsymbol{v}$ 和聚类中心 C。

隶属度矩阵 $\boldsymbol{v}_{ij}=\dfrac{1}{\sum_{p=1}^{k}\left(\dfrac{\|X_i-C_j\|}{\|X_i-C_P\|}\right)^{\frac{2}{b-1}}}$ (5)

聚类中心 $C_j=\dfrac{\sum_{i=1}^{n}\boldsymbol{v}_{ij}^{b}\dfrac{X_i\rho_i}{\bar{\rho}}}{\sum_{i=1}^{n}\boldsymbol{v}_{ij}^{b}}$ (6)

(4)若 $|J^{(l)}-J^{(l-1)}|\leqslant\varepsilon$,则结束;否则返回步骤(3)继续迭代。

目标函数 $\min\ J_{RCM}(\boldsymbol{v},C)=\sum_{i=1}^{m}\sum_{j=1}^{k}V_j\left(\dfrac{X_i\rho_i}{\bar{\rho}}\right)^{b}\left\|\dfrac{X_i\rho_i}{\bar{\rho}}-C_j\right\|^2$

约束条件 $s.t.\ \sum_{j=1}^{k}\boldsymbol{v}_i\left(\dfrac{X_i\rho_i}{\bar{\rho}}\right)=1,0<\sum_{j=1}^{n}\boldsymbol{v}_j\left(\dfrac{X_i\rho_i}{\bar{\rho}}\right)<n$

3 实验

本文采集和提取局域网数据特征值:{网络协议,流量,连接频率}[4],进行实验主要的分析方法为上文所提的改进的 FCM 算法。

图 1 是经归一化处理的实验数据,其中 * 为经预处理后的初始化聚类中心,相应的坐标为:C_1(0.062 4,0.034 8,0.075 2),C_2(0.075 4,0.519 2,0.084 6),C_3(0.093 1,0.965 3,0.093 4)

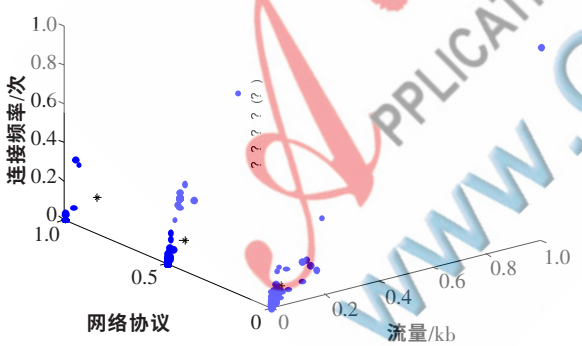

图 1 实验数据

实验结果如下:

(1)将改进的 FCM 算法与原始的 FCM 算法在时间效率上进行对比。如图 2、图 3 所示。

(2)在聚类效果上进行对比,如表 1 所示。

在表格中,Y_i 表示第 i 类包含的元素个数,S_i 表示实验所检测到的第 i 类所包含的元素个数,而 $Y_i\cap S_i$ 为真正属于第 i 类的元素个数。$Y_i\cap S_i/S_i$ 为聚类的精度(Precision),$Y_i\cap S_i/Y_i$ 为聚类的召回率(Recall)。

从实验结果可以看出:

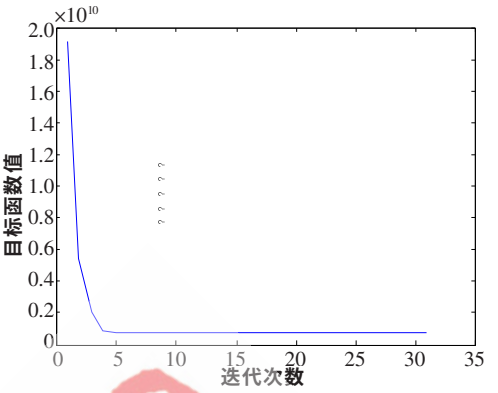

图 2 改进后的 FCM 目标函数趋势图

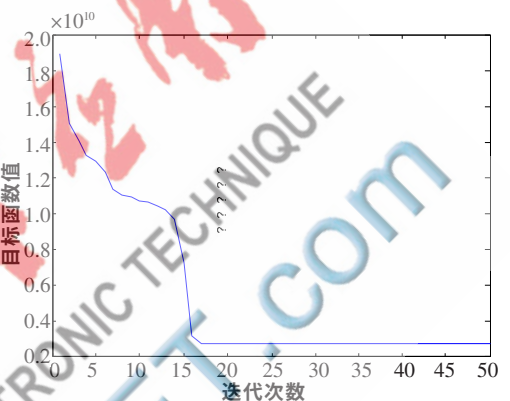

图 3 原始的 FCM 目标函数趋势图

表 1 聚类效果的实验结果

算法	原始 FCM			改进 FCM		
	i=1	i=2	i=3	i=1	i=2	i=3
Y_i	20	90	200	20	90	200
S_i	20	106	184	20	96	194
$Y_i\cap S_i$	20	75	159	20	80	174
Recall(%)	100%	83%	79.5%	100%	88%	87%
Precision(%)	100%	70.8%	86.4%	100%	83.3%	89.6%

(1)在改进的 FCM 算法中,计算过程只迭代了 31 次出结果。而原始的 FCM 算法迭代了 74 次,才出结果。在时间效率上,改进的 FCM 算法有明显的提高。

(2)改进的 FCM 算法在迭代过程中比较平稳,并且收敛的速度也比较快。然而未改进的 FCM 算法虽然总的趋势也是收敛的,可是收敛的速度存在一定的波动。

本文针对 FCM 算法中存在的计算时耗高,对初始化数据敏感,容易陷入局部最优的问题,引入网格聚类,加权计算,寻找连通分量,自适应免疫算法等方法并采用初始化预处理的方法,对其进行一定的改进。在试验中采集局域网中的数据包使实验数据随机分布,存在局部最优解。与原始的 FCM 算法进行对比试验,结果表明,改进的 FCM 算法在时间效率和收敛速度上都有明显的提高,改进的 FCM 算法是有效的。

参考文献

[1] XINCHAO Z. Simulated annealing algorithm with adaptive neighborhood [J]. Applied Soft Computing, 2011,11(2): 1827-1836.

[2] SZABO A, CASTRO L N D, DELGADO M R. FaiNet: An immune algorithm for fuzzy clustering [C]. in Fuzzy Systems (FUZZ-IEEE). 2012 IEEE International Conference on. 2012.

[3] HATHAWAY R J, HU Y. Density-weighted fuzzy c-means clustering[J]. IEEE Transactions on Fuzzy Systems, 2009. 17(1): 243-252.

[4] XING W, ZHAO Y, LI T. Research on the defense against ARP spoofing attacks based on Winpcap [C]. in Education Technology and Computer Science (ETCS), 2010 Second International Workshop on, 2010.

（收稿日期：2013-09-23）

作者简介：

丁旭晨，男，1987 年生，硕士研究生，主要研究方向：物联网、信息安全。

林锦贤，男，1957 年生，教授，博士，主要研究方向：计算机网络、数据库、信息安全。